# LINEAIRE REGRESSIE: DE BASIS

14 mei 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)

# WIE BEN IK?

Elmar Jansen (elmar@elmarjansen.nl)

Docent Sociale Wetenschappen (Universiteit van Amsterdam)

- methoden- en statistiekonderwijs
- diverse vakken sociaal wetenschappelijke theorie

Algehele statistiek-, data- en programmeernerd

# VANDAAG

1. Regressie: de basis
   a) Waarom regressie?
   b) Wat is regressie?
2. Enkelvoudige Lineaire Regressie
   a) Conceptueel: regressievergelijking en grafiek
   b) Interpretatie van coëfficiënten
3. De $R^2$
4. Omgaan met onzekerheid
   a) Standaardfout
   b) T-toets(en)
   c) F-toets
5. Meervoudige Regressie
   a) Conceptueel: regressievergelijking
   b) Interpretatie van coëfficiënten
6. Afsluiting

# DE KOMENDE WEKEN

| Bijeenkomst | Onderwerp |
|---|---|
| Dinsdag 14 mei | Lineaire regressie: de basis |
| Dinsdag 21 mei | Lineaire regressie vervolg: assumpties, controleren en dummy-variabelen |
| Donderdag 30 mei | Logistische Regressie |
| Dinsdag 4 juni | Interacties |
| Dinsdag 11 juni | Multilevel-analyse |

# WAAROM REGRESSIE?

# RELATIES TUSSEN VARIABELEN

- In statistisch onderzoek willen we vaak meer dan alleen losse variabelen samenvatten: we zijn geïnteresseerd in **verbanden** en **relaties** tussen verschillende variabelen

- Als die variabelen gemeten zijn op **interval/ratio-niveau** (ze zijn numeriek), is regressie vaak de beste manier om deze relaties te onderzoeken

# HOE ZAT HET OOK ALWEER MET MEETNIVEAUS?

**KWALITATIEF**
"overige opmerkingen"

Categoriaal

**DICHOTOOM**
gepensioneerd?

+ meer dan 2 categorieën

**NOMINAAL**
partijkeuze

+ volgorde

**ORDINAAL**
opleidingsniveau

+ gelijke afstanden

Numeriek

**INTERVAL**
temperatuur in C

+ betekenisvol nulpunt

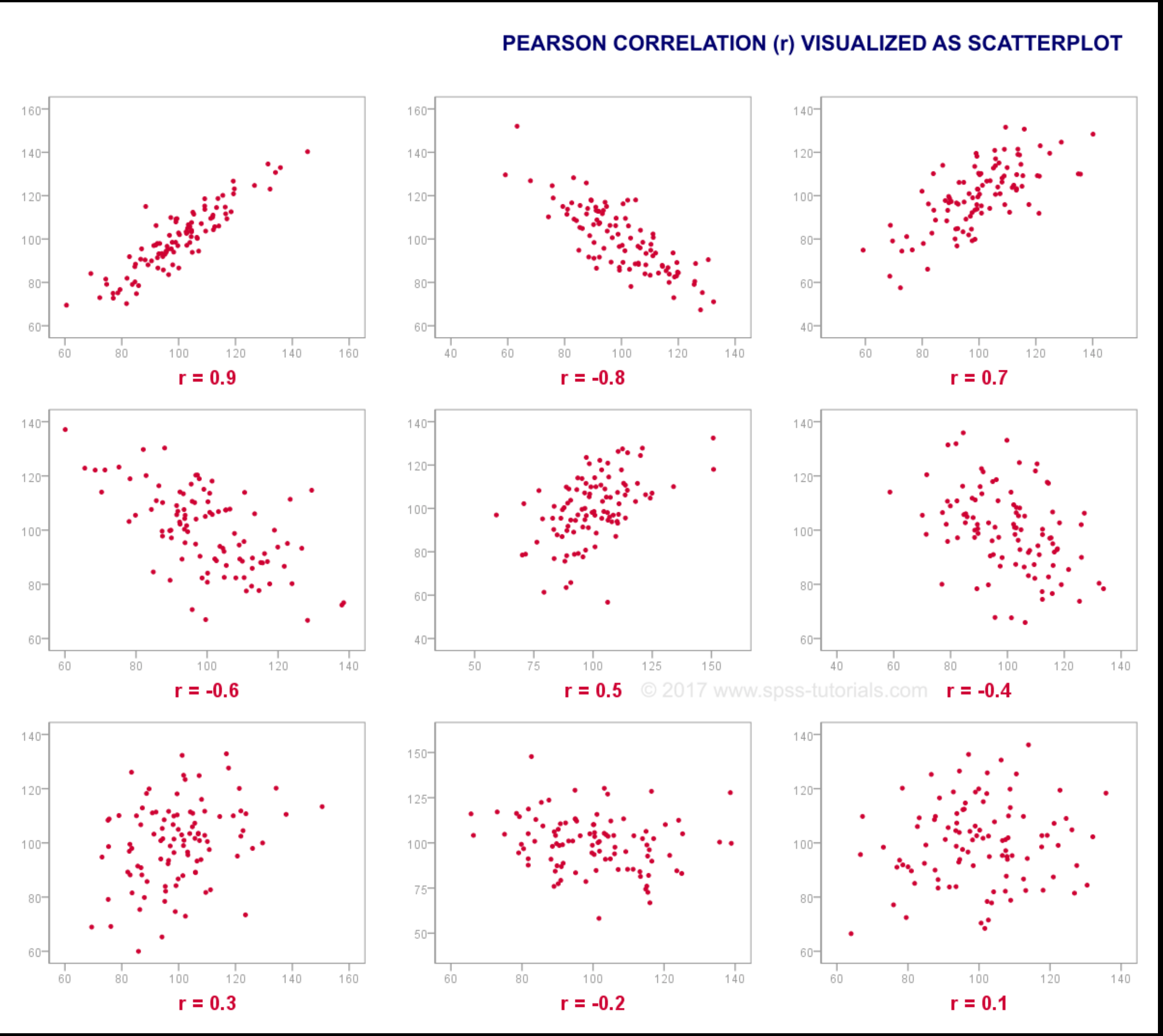
**RATIO**
gewicht in kg

# HOE ZAT HET OOK ALWEER MET PEARSONS CORRELATIE COËFFICIËNT *R?*

Pearsons r is:

- Maat voor samenhang tussen twee numerieke variabelen
- Getal tussen -1 en 1

Dus: we hebben al een eenvoudige maat om relatie tussen twee interval-variabelen uit te drukken…



# WAAROM REGRESSIE?

- Gaat uit van **causaliteit** en is **asymmetrisch** met **onafhankelijke variabelen** (de “oorzaak”) en één **afhankelijke variabele** (het “gevolg”).
- Om een **model van de werkelijkheid** te maken waarmee je kunt **voorspellen**
- Om **nauwkeurige en concrete uitspraken** te doen over het **effect** van de ene variabele op de andere
- Met meervoudige regressie kun je **controleren** voor (o.a.) schijnverbanden
- Het is **veelzijdig**: biedt een eenvoudig model dat (bijna) alle andere statistische toetsen kan vervangen

# NADELEN REGRESSIE?

- Lineaire regressie werkt alleen voor **lineaire verbanden**
- Er mag geen sprake zijn van een **schijnverband**
- Vrij veel **technische voorwaarden** waaraan moet zijn voldaan (zie volgende week)
- Makkelijk om **verkeerde conclusies** te trekken
  - Veel mogelijke problemen zonder eenvoudige (technische) oplossing. Je moet dus goed weten waar je mee bezig bent
  - Interpretatie vergt enige oefening

# ENKELVOUDIGE REGRESSIE

Lineaire regressie met één onafhankelijke variabele

# LINEAIRE REGRESSIE

Maakt een (lineair) model

om

de waarden te voorspellen

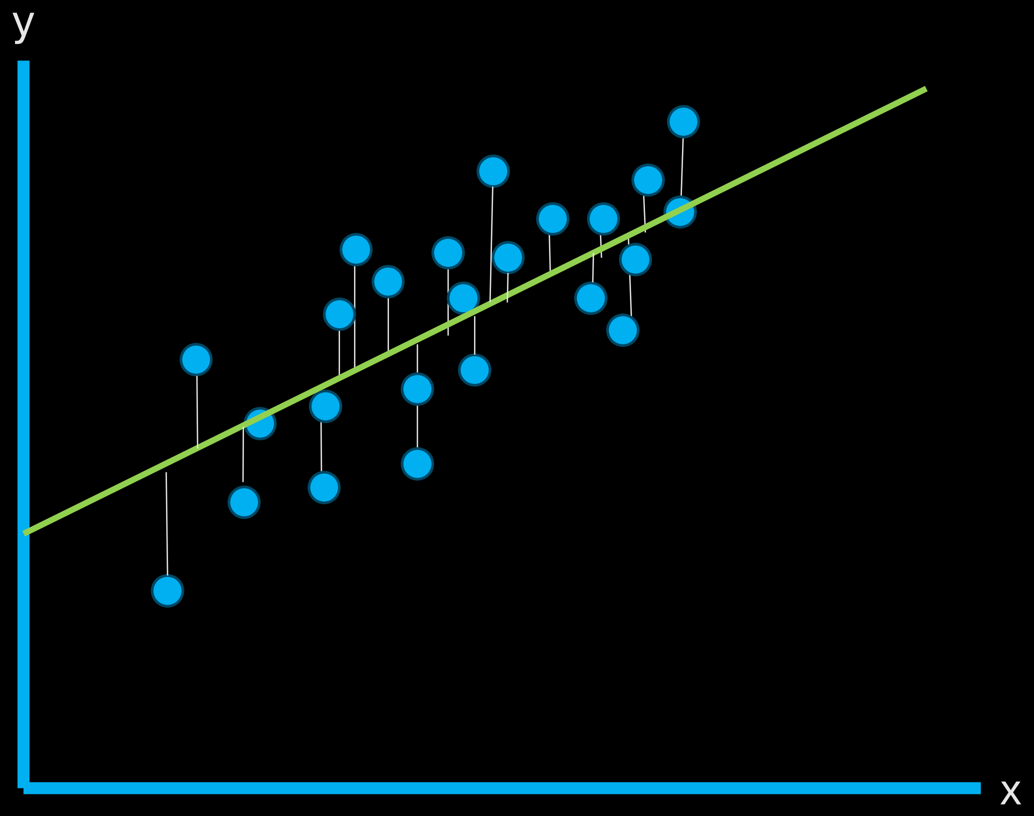
van een *afhankelijke variable* y

met de waarden van

één of meer

*onafhankelijke variabelen* x

# LINEAIRE REGRESSIE IN GRAFIEK




Als we een regressie “draaien”,
vragen we de software om
een lijn te trekken,
zo dat de verticale afstanden tussen
de punten en de lijn minimal zijn

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

# LINEAIRE REGRESSIE IN VERGELIJKING

$$\boxed{y_i} = \beta_0 + \beta_1 \boxed{x_i} + \boxed{\varepsilon_i}$$

Afhankelijke variabele (waarde voor eenheid i)

Onafhankelijke variabele (waarde voor eenheid i)

Een onverklaarde afwijking (residu) (voor eenheid i) ☹

# LINEAIRE REGRESSIE IN VERGELIJKING

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Als we een regressie “draaien”,
vragen we de software om
een *schatting* van $\beta_0$ en $\beta_1$ te vinden
zo dat de residuen ($\varepsilon_i$)
zo klein mogelijk zijn

Of preciezer: we minimaliseren $\sum(\varepsilon_i)^2$.
Vandaar: Ordinary Least Squares of OLS

# LINEAIRE REGRESSIE IN VERGELIJKING

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Griekse letters omdat de *coefficienten* hier *populatie*-parameters zijn.

We *weten* niet wat de waarden van deze parameters zijn, maar we kunnen ze wel *schatten* en er *hypothesen* over formuleren.

# LINEAIRE REGRESSIE IN FORMULE

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

alternatieve notaties:

$\hat{y}_i = \beta_0 + \beta_1 x_i$ (*voorspelde waarde* of *predicted value*)

$E(y_i) = \beta_0 + \beta_1 x_i$ (*verwachtingswaarde* of *expected value*)

$$gewicht_i = \beta_0 + \beta_1 lengte_i + \varepsilon_i$$

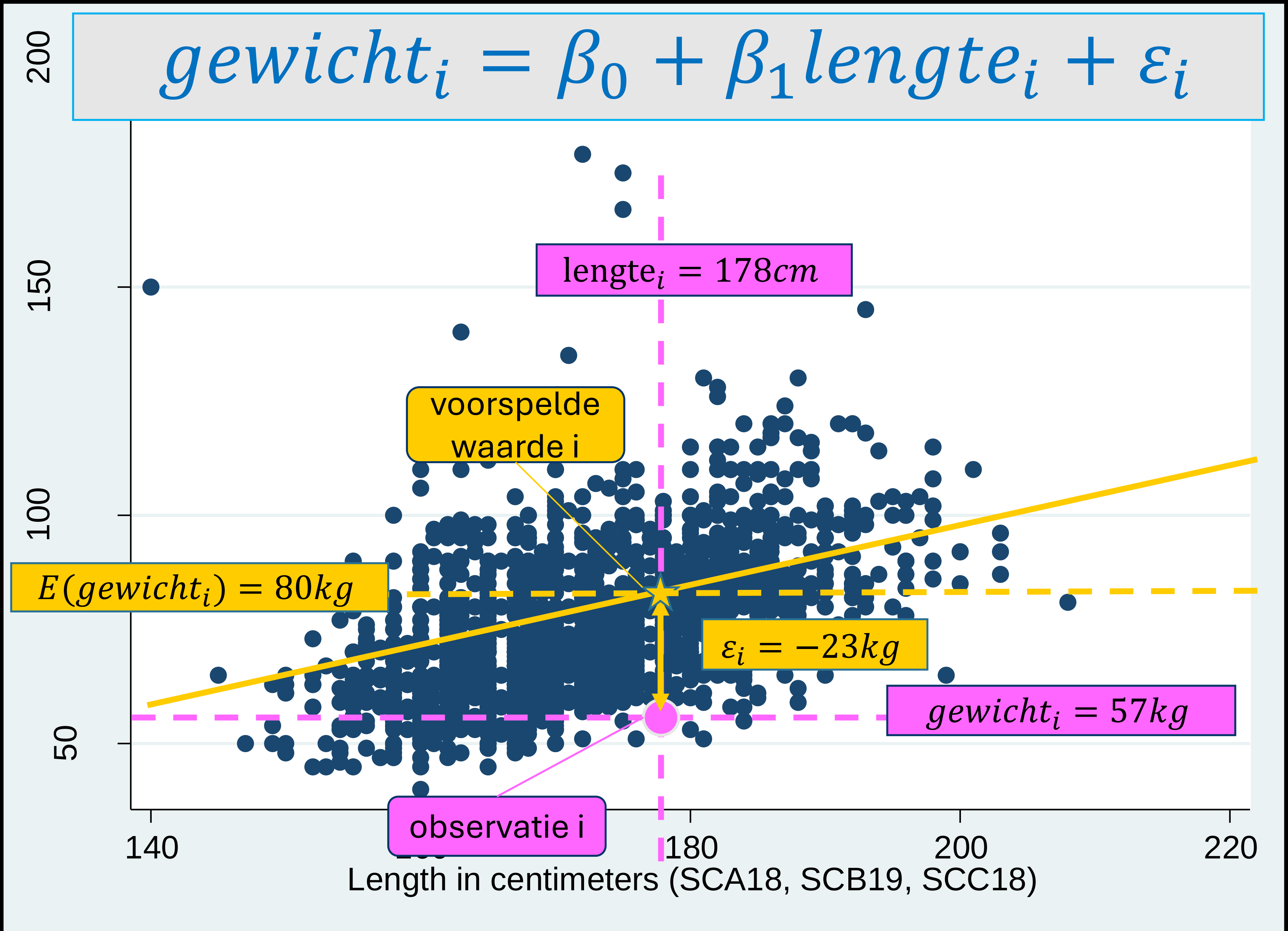
# LINEAIRE REGRESSIE

**Voorbeeld:**

Voorspel lichaamsgewicht op basis van lengte

# VOORBEELD IN GRAFIEK

$$gewicht_i = \beta_0 + \beta_1 lengte_i + \varepsilon_i$$



# VOORBEELD IN R

```
> model <- lm(gewicht ~ lengte, data=dt)
> summary(model)

Call:
lm(formula = gewicht ~ lengte, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.169  -8.716  -1.600   7.124 107.124 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -59.11937    5.89162  -10.03   <2e-16 ***
lengte        0.78450    0.03384   23.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.99 on 1761 degrees of freedom
Multiple R-squared:  0.2338,	Adjusted R-squared:  0.2334 
F-statistic: 537.5 on 1 and 1761 DF,  p-value: < 2.2e-16
```

$$gewicht_i = \boxed{\beta_0} + \boxed{\beta_1} lengte_i + \varepsilon_i$$

$$gewicht_i = -59.1 + 0.785 \times lengte_i + \varepsilon_i$$

# LINEAIRE REGRESSIE: RESULTATEN

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$gewicht_i = -59.1 + 0.785 \times lengte_i + \varepsilon_i$$

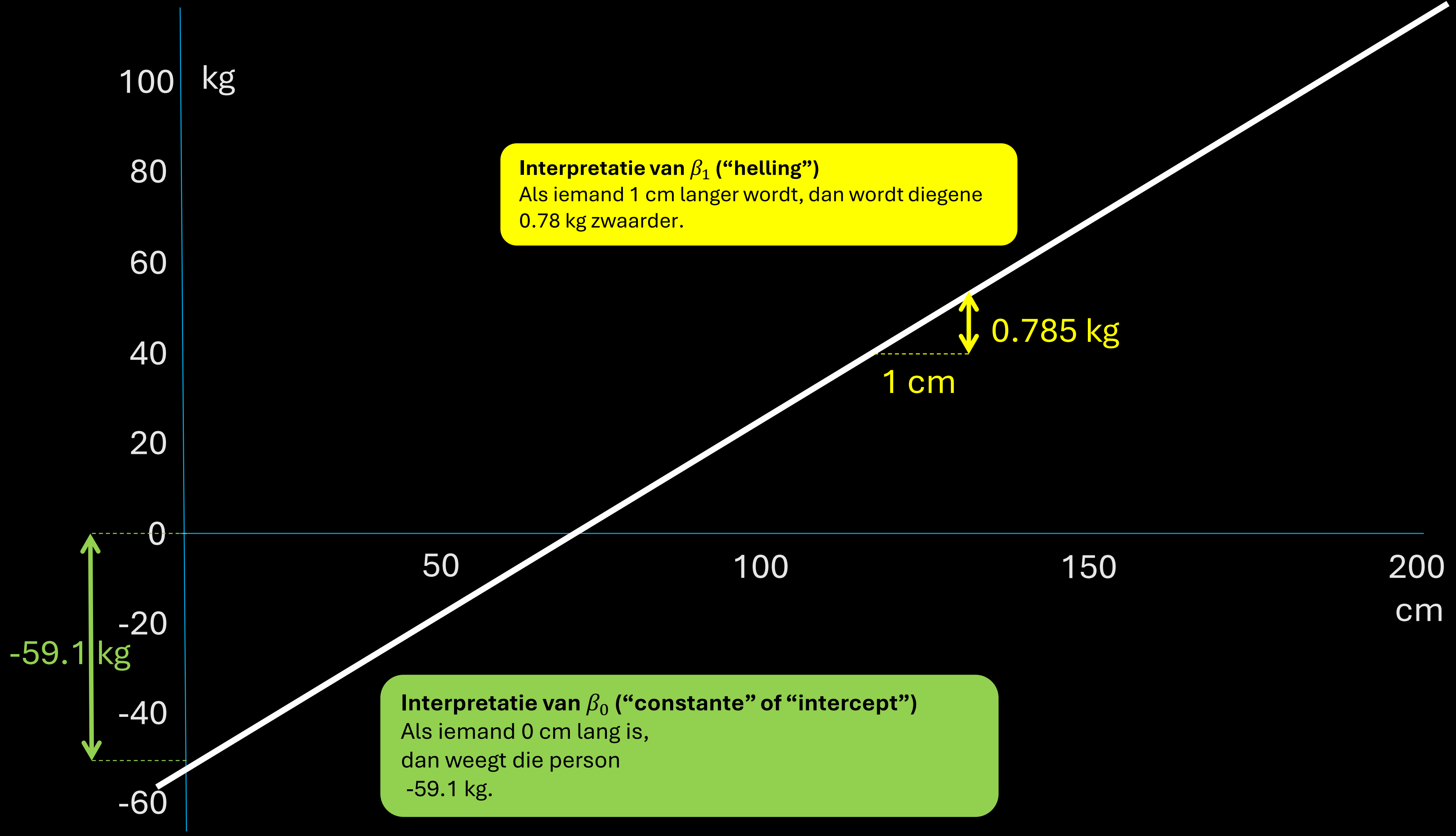
**Interpretatie van $\beta_0$ ("constante" of "intercept")**
Als iemand 0 cm lang is, dan weegt die person -59.1 kg.

**Interpretatie van $\beta_1$ ("helling")**
Als iemand 1 cm langer wordt, dan wordt diegene 0.78 kg zwaarder.

# IN GRAFIEK

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$gewicht_i = -54.8 + 0.756 \times lengte_i + \varepsilon_i$$



https://elmarjansen.nl/os

# OEFENING 1

Het effect van gemiddelde grootte van huishoudens in een buurt
op gemiddeld aantal auto's per huishouden in die buurt

# RESULTATEN OEFENING

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35595 -0.12057 -0.06430  0.07108  1.81900

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.34603    0.05040  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.02729  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,657[a] | ,432 | ,431 | ,2057 |

a. Predictors: (Constant), Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 13,033 | 1 | 13,033 | 308,156 | <.001[b] |
| | Residual | 17,128 | 405 | ,042 | | |
| | Total | 30,161 | 406 | | | |

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

b. Predictors: (Constant), Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | -,346 | ,050 | | -6,866 | <.001 | -,445 | -,247 |
| | Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal) | ,479 | ,027 | ,657 | 17,554 | <.001 | ,425 | ,533 |

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35595 -0.12057 -0.06430  0.07108  1.81900 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.34603    0.05040  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.02729  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307 
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

**Model Summary**

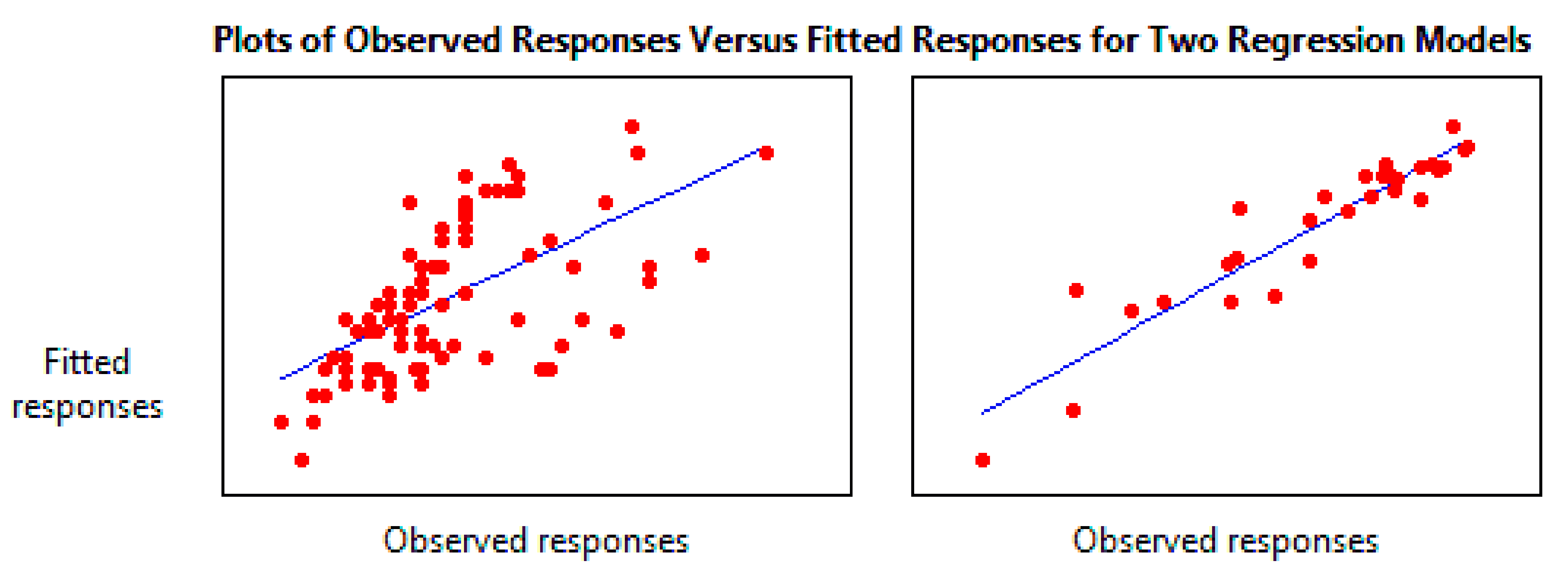
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,657[a] | ,432 | ,431 | ,2057 |

a. Predictors: (Constant), Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)

# DE $R^2$

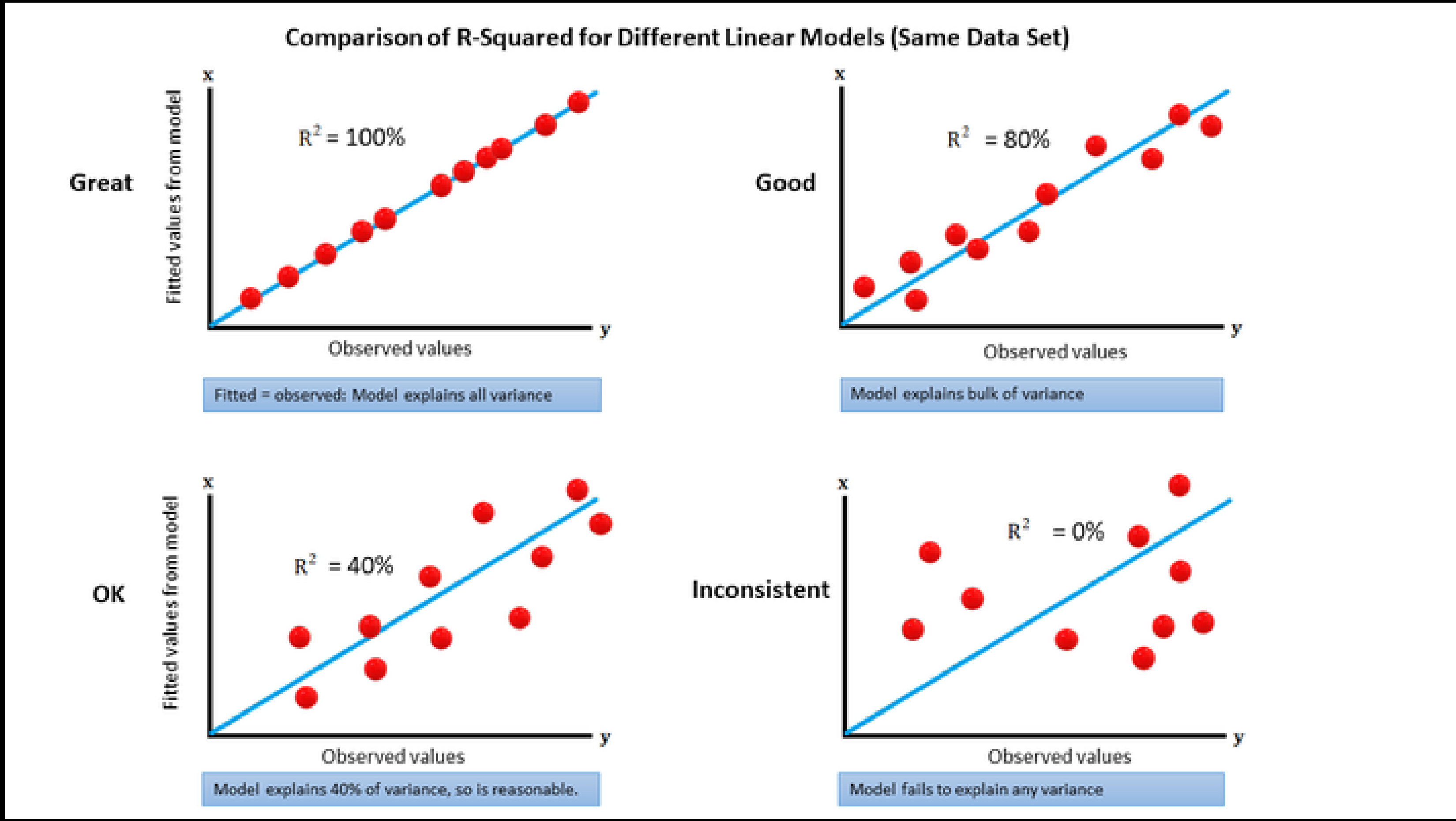
Maat voor kwaliteit van het model, oftewel: *model fit*

# HOE KUNNEN WE ONDERSCHEID MAKEN?



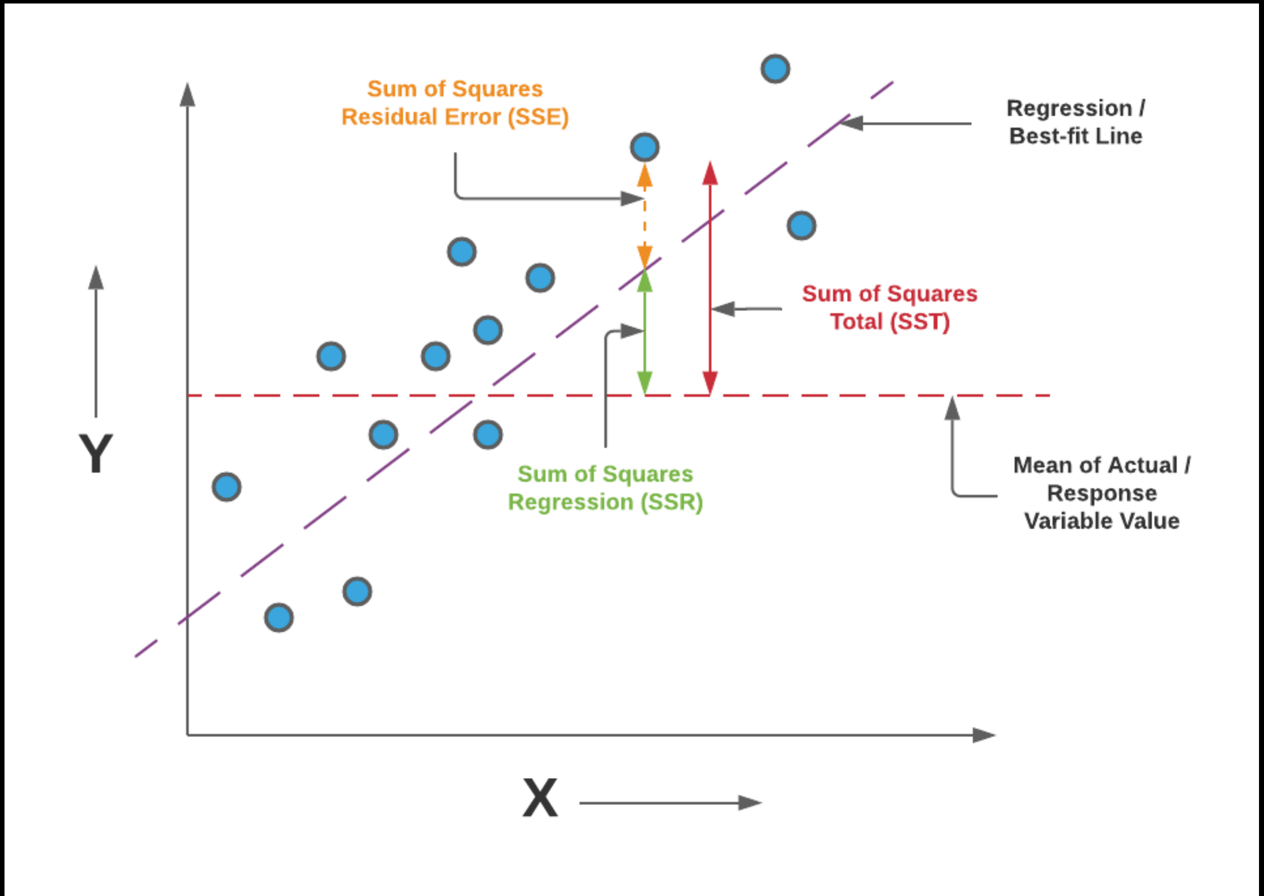
# DE $R^2$

De $R^2$ geeft de ***verklaarde variantie***: een indicatie van hoe goed de gemaakte vergelijking (het "model") de afhankelijke variabele voorspelt.



# R-KWADRAAT BEREKENEN

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$



# DE ADJUSTED $R^2$

De Adjusted $R^2$ is een iets **gecorrigeerde $R^2$** .

Hij weegt mee dat je model **altijd beter gaat voorspellen** *in je steekproef* als je onafhankelijke **variabelen toevoegt** aan het model.

Dit is vooral relevant **bij kleine N** en **veel onafhankelijke variabelen**.

# $R^2$

$$weight_i = -54.8 + 0.756 \times height_i + \varepsilon_i$$

```
Call:
lm(formula = gewicht ~ lengte, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.169  -8.716  -1.600   7.124 107.124 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -59.11937    5.89162  -10.03   <2e-16 ***
lengte        0.78450    0.03384   23.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.99 on 1761 degrees of freedom
Multiple R-squared:  0.2338,	Adjusted R-squared:  0.2334
```

De $R^2$ geeft de ***verklaarde variantie***: een indicatie van hoe goed de gemaakte vergelijking (het "model") de afhankelijke variabele voorspelt. Onze op lengte gebaseerde vergelijk kan 23% van alle variatie in gewicht verklaren.

# $R^2$

Hoe zit het in het model dat we net zelf hebben gedraaid?

De $R^2$ van ons aantal-auto’s-model is 0.43.

De verschillen in het aantal auto’s per huishouden tussen buurten kunnen dus voor 43% verklaard worden door verschillen in (gemiddelde) huishoudomvang.

# OMGAAN MET ONZEKERHEID

Van steekproef naar populatie:

manieren om *onzekerheid* van onze *schattingen* uit te drukken

# INFERENTIËLE STATISTIEK

Op basis van een bekende

(maar oninteressante) **steekproef**

schattingen doen

ten aanzien van de **populatie**.

# STANDAARDFOUT

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35595 -0.12057 -0.06430  0.07108  1.81900 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.34603    0.05040  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.02729  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321,	Adjusted R-squared:  0.4307 
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

**Formeel:** de geschatte standaardafwijking van de *steekproevenverdeling* van de geschatte parameter

**Intuïtief:** een indicatie van hoe ver we denken dat de schatting (gemiddeld) van de echte waarde af zit.

We verwachten dus dat het effect van hh_grootte 0.48 is,
maar met deze steekproef zitten we daar gemiddeld 0.03 naast.

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | -,346 | ,050 | | -6,866 | <.001 | -,445 | -,247 |
| | Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal) | ,479 | ,027 | ,657 | 17,554 | <.001 | ,425 | ,533 |

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

# T-TOETS (SIGNIFICANTIE VAN COËFFICIËNTEN)

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35595 -0.12057 -0.06430  0.07108  1.81900 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.34603    0.0504   -6.866 2.49e-11 ***
hh_grootte   0.47912    0.0272   17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307 
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

**Toets** of de gevonden steekproef waarschijnlijk is als de coëfficiënt in de populatie eigenlijk 0 is.

**Met andere woorden:**
zou je dit effect toevallig kunnen vinden in een steekproef, als er eigenlijk geen effect is.

$p < 0.01$ voor het effect van hh_grootte:

Het effect is significant: het is niet waarschijnlijk om deze steekproef te vinden als er in de populatie geen effect is.

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95.0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound |
| 1 | (Constant) | -,346 | ,050 | | -6,866 | <.001 | -,445 | -,247 |
| | Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal) | ,479 | ,027 | ,657 | 17,554 | <.001 | ,425 | ,533 |

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

Let op: de significantie van de constante is inhoudelijk niet interessant

# F-TOETS (SIGNIFICANTIE VAN MODEL)

```
Call:
lm(formula = autos_per_hh ~ hh_grootte, data = buurten)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35595 -0.12057 -0.06430  0.07108  1.81900 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.34603    0.05040  -6.866 2.49e-11 ***
hh_grootte   0.47912    0.02729  17.554  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2057 on 405 degrees of freedom
  (72 observations deleted due to missingness)
Multiple R-squared:  0.4321, Adjusted R-squared:  0.4307 
F-statistic: 308.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

**ANOVA**[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 13,033 | 1 | 13,033 | 308,156 | <.001[b] |
| | Residual | 17,128 | 405 | ,042 | | |
| | Total | 30,161 | 406 | | | |

a. Dependent Variable: Motorvoertuigen/Personenauto's/Personenauto's per huishouden (per huishouden)

b. Predictors: (Constant), Bevolking/Particuliere huishoudens/Gemiddelde huishoudensgrootte (aantal)

**Toets** of de gevonden verklaringskracht van het model (de R2) waarschijnlijk is als het model in werkelijkheid geen enkele voorspellende kracht heeft.

**Met andere woorden:**
zou je dit gehele model kunnen vinden in een steekproef, als eigenlijk geen onafhankelijke variabele effect heeft.

$p < 0.01$:

Het model voorspelt significant beter dan geen model

Let op: dit is eigenlijk alleen relevant bij meerdere onafhankelijke variabelen.

# MEERVOUDIGE REGRESSIE

Regressie met meerdere onafhankelijke variabelen

# MEERVOUDIGE REGRESSIE

Meervoudige Regressie

Regressie met meerdere onafhankelijke variabelen

$$\boxed{y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \varepsilon_i}$$

**Interpretatie van $\beta_1$**

Als $x_1$ omhoog gaat met 1
**en $x_2$ gelijk blijft**, stijgt y met $\beta_1$

**Interpretatie van $\beta_2$**

Als $x_2$ omhoog gaat met 1
**en $x_1$ gelijk blijft**, stijgt y met $\beta_2$

Effect is nu *constant houdend voor* andere variabele
Dat gaan we volgende week gebruiken om te *controleren*

# MEERVOUDIGE REGRESSIE

*Verder verandert er niets ten opzichte*

*van enkelvoudige regressie!*

# OEFENING 2

Gemiddeld aantal auto's per huishouden verklaard op basis van huishoudengrootte en koopwoningen

# AFSLUITING

# VOLGENDE WEEK

- Controleren en causaliteit
- Regressie-assumpties
- Transformaties
- Dummy-variabelen
- Rapporteren

# TOT SLOT

- De slides zijn online beschikbaar (https://elmarjansen.nl/os)
- Een handout / reader met de samenvatting is in de maak
- Deel je opmerkingen / wensen / vragen!
  elmar@elmarjansen.nl